Predictors of individual performance changes related to item positions in PISA assessments

Wu, Qian ; Debeer, Dries ; Buchholz, Janine ; Hartig, Johannes ; Janssen, Rianne

Abstract: Background: Item position effects have been a common concern in large-scale assessments as changing the order of items in booklets may have an undesired effect on test performance. If every test taker would be affected by the effect in the very same way, comparisons between groups of individuals would still be valid. However, research has shown that in addition to a general fixed effect of item positions, the extent of the effect varies considerably across individuals. These individual differences are referred to as persistence. Test takers with a high level of persistence are able to keep up their performance better throughout the test administration, whereas those with a lower level of persistence show a larger decline in their test performance. Methods: The present study applied a multilevel extended item response theory (IRT) framework and used the data from the PISA 2006 science, 2009 reading, and 2012 mathematics assessments. The first objective of this study is to provide a systematic investigation of item position effects across the three PISA domains, partially replicating the previous studies on PISA 2006 and 2009. Second, this study aims to gain a better understanding of the nature of individual differences in position effects by relating them to student characteristics. Gender, socio-economic status, language spoken at home, and three motivational scales (enjoyment of doing the subject being assessed, effort thermometer, perseverance) were used as person covariates for persistence. Results: This study replicated and extended the results found in previous studies. An overall negative item cluster position effect and significant individual differences in this effect were found in all the countries in the three PISA domains. Furthermore, the most frequently observed effect of person covariates on persistence is gender, with girls keeping up their performance better than boys. Other predictors showed little or inconsistent effects on persistence. Conclusions: Our study demonstrated inter-individual differences as well as group differences in item position effects, which may threaten the comparability between persons and groups. The consequences and implications of item position effects and persistence for the interpretation of PISA results are discussed.

DOI: https://doi.org/10.1186/s40536-019-0073-6

Originally published at:
Wu, Qian; Debeer, Dries; Buchholz, Janine; Hartig, Johannes; Janssen, Rianne (2019). Predictors of individual performance changes related to item positions in PISA assessments. Large-Scale Assessments in Education, 7(1):5.
DOI: https://doi.org/10.1186/s40536-019-0073-6

Wu *et al. Large-scale Assess Educ* (2019) 7:5
https://doi.org/10.1186/s40536-019-0073-6

Large-scale Assessments in Education

RESEARCH Open Access

Predictors of individual performance changes related to item positions in PISA assessments

Qian Wu[1*], Dries Debeer[2], Janine Buchholz[3], Johannes Hartig[3] and Rianne Janssen[1]

*Correspondence:
qian.wu@kuleuven.be
[1] KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

Abstract

Background: Item position effects have been a common concern in large-scale assessments as changing the order of items in booklets may have an undesired effect on test performance. If every test taker would be affected by the effect in the very same way, comparisons between groups of individuals would still be valid. However, research has shown that in addition to a general fixed effect of item positions, the extent of the effect varies considerably across individuals. These individual differences are referred to as persistence. Test takers with a high level of persistence are able to keep up their performance better throughout the test administration, whereas those with a lower level of persistence show a larger decline in their test performance.

Methods: The present study applied a multilevel extended item response theory (IRT) framework and used the data from the PISA 2006 science, 2009 reading, and 2012 mathematics assessments. The first objective of this study is to provide a systematic investigation of item position effects across the three PISA domains, partially replicating the previous studies on PISA 2006 and 2009. Second, this study aims to gain a better understanding of the nature of individual differences in position effects by relating them to student characteristics. Gender, socio-economic status, language spoken at home, and three motivational scales (enjoyment of doing the subject being assessed, effort thermometer, perseverance) were used as person covariates for persistence.

Results: This study replicated and extended the results found in previous studies. An overall negative item cluster position effect and significant individual differences in this effect were found in all the countries in the three PISA domains. Furthermore, the most frequently observed effect of person covariates on persistence is gender, with girls keeping up their performance better than boys. Other predictors showed little or inconsistent effects on persistence.

Conclusions: Our study demonstrated inter-individual differences as well as group differences in item position effects, which may threaten the comparability between persons and groups. The consequences and implications of item position effects and persistence for the interpretation of PISA results are discussed.

Keywords: Item position effects, Persistence, Multilevel IRT, PISA

Introduction

In large-scale assessments, items are often presented in different positions within a test using a booklet design. However, changes in the presentation order of items can have unintended effects on item characteristics and hence on test takers' performance (Leary and Dorans 1985). That is, when presented at the end of a test, an item may become less or more difficult than when presented at the beginning of the test, resulting in an increase or a decrease in the performance during testing, respectively. Under such circumstances, the responses to test items are not solely dependent on the latent trait(s) of interest, but also on context factors that may be construct-irrelevant (Messick 1995). This violates the assumption of local independence of the item response theory (IRT) models that are commonly used to analyze assessment data, and would bias the item and person parameter estimates, invalidating the interpretation of individuals' test results (AERA, APA, and NCME 2014).

A number of studies have investigated the effect of item positions in different content domains using distinct methodological approaches. Two types of position effects have been reported. On the one hand, a positive item position effect (i.e., an item becomes less difficult when administered in the later position of a test) can occur because of a learning or practice effect when test takers become more acquainted with the test material, format, and the kind of tasks asked. Kingston and Dorans (1984) compared mean differences in the IRT difficulty parameters of the Graduate Record Examination (GRE) test items and found a large and consistent practice effect for item types of analysis of explanations and logical diagrams. Verguts and De Boeck (2000) developed a dynamic Rasch model and showed a learning effect in a computer-administered intelligence test. Schweizer et al. (2009) used confirmatory factor analysis and reported a position dimension that might represent learning for the Advanced Progressive Matrices (APM) items.

On the other hand, a negative item position (i.e., an item becomes more difficult when administered in the later position) has been reported more frequently. Kingston and Dorans (1984) found a moderate increase in item difficulties for items of reading comprehension in the same GRE test. Hohensinn et al. (2008) found a small negative item position effect in a 4th grade math test by means of the linear logistic test model. Meyers et al. (2009) compared the field-testing and operational testing of standardized tests of reading and mathematics, and showed that Rasch item difficulty increased with changes in item positions. Hartig and Buchholz (2012) employed a multilevel IRT model and reported a significant negative item position effect in 10 selected countries in the Programme of International Student Achievement (PISA) 2006 science assessment. Debeer et al. (2014) used the same approach and found a negative position effect in all countries in the PISA 2009 reading assessment. This negative effect has been commonly attributed to fatigue or declining motivation of test takers during testing. Weirich et al. (2016) demonstrated that position effects were only partially due to changes in test-taking effort and did not vanish when the (decline in) test-taking effort was controlled for. Other context factors may also contribute to position effects, such as testing time and speed, as was the case for the Trends in International Mathematics and Science Study (TIMSS) 2003 (Mullis et al. 2004).

If every test taker would be affected by these effects in the very same way, ignoring position effects would still lead to fair comparisons. However, if the strength of item

position effects varies across persons, then the unadjusted test scores may lead to biased ability parameter estimates. In fact, recent research shows that in addition to a general effect of item positions, there can be substantial individual differences regarding this effect (Debeer and Janssen 2013; Schweizer et al. 2009, 2011; Verguts and De Boeck 2000). Wise et al. (1989) found that the variation in performance associated with item positions was more related to the ability of examinees than to the characteristics of items, and the effect was more pronounced for lower ability examinees. Significant individual differences regarding position effects were also found in the PISA 2006 science and 2009 reading assessments, with more variation in lower performing countries (Debeer et al. 2014; Hartig and Buchholz 2012). Weirich et al. (2016) further found that position effects can be moderated by the decline in test-taking effort. Given the more frequently observed negative position effect in large-scale assessments, the term "persistence" is used to refer to these individual differences, describing test takers' capability of keeping up their performance during testing sessions (Hartig and Buchholz 2012). Low individual persistence indicates a decline in test performance, while high persistence means a maintained or even increased performance level during testing.

The present study

To summarize, previous research employed different methodological approaches to investigate item position effects in assessments of different natures (e.g., low-/high-stakes, summative, or problem-solving). The results were not always consistent, depending on the content domains, item formats, and test administration contexts. Thus, it would be intriguing to examine the effects of item positions within a consistent methodological framework for assessments with varying content domains but with similar test administration procedures. Therefore, the first objective of the present study is to provide a systematic investigation of item position effects in large-scale assessments by analyzing the three major domains of the PISA 2006, 2009, and 2012 assessments (i.e., science, reading, and mathematics, respectively). More specifically, the study investigates (1a) the general effect of item positions, (1b) persistence, and (1c) the correlations of persistence with ability in the three domains.

The second objective is to gain a better understanding of persistence by providing an exploratory investigation relating those individual differences to student characteristics. Until now, relatively little research has explicitly addressed the question why item position effects occur and which personal variables can explain differences in persistence. The limited findings in the literature suggest that persistence is likely related to the ability (Wise et al. 1989) and to the test-taking effort (Weirich et al. 2016) of students. Therefore, this study investigates the relationship of persistence in PISA assessments with (a) student background characteristics that have been known as stable predictors of test performance (e.g., gender, socio-economic status, and language spoken at home) and (b) motivational variables that have been included in the PISA context questionnaires (enjoyment of the subject being assessed, "effort thermometer", and self-reported perseverance). With these exploratory analyses, more insights may be obtained into (2a) which specific subgroups of students are more prone to lower persistence and (2b) whether the individual variation of the effect is due to individual differences in the general interest of the test subject and/or the test-taking effort.

We follow the multilevel IRT modeling approach proposed by Hartig and Buchholz (2012) and Debeer and Janssen (2013). The choice of this approach has three reasons. Firstly, IRT is the most widely used theoretical framework when it comes to large-scale assessments (Berezner and Adams 2017). Secondly, using the same modeling approach, Hartig and Buchholz (2012) studied item position effects in 10 selected countries of the PISA 2006 science assessment, and Debeer et al. (2014) analyzed all the participating countries of the PISA 2009 reading assessment. Thus, the current study partly serves as a replication of the previous two studies, but also extends them by studying all the participating countries and including the third major domain of PISA assessments, mathematics in PISA 2012. Thirdly and more importantly, this multilevel extension of the IRT model enables a closer examination into the nature of individual differences in item position effects by incorporating explanatory person covariates into the model (see below).

Modeling approach

Within the framework of the generalized linear mixed approach (GLMM) to IRT models (De Boeck and Wilson 2004), the position of an item can be specified as a predictor to investigate its effect on the probability of success on the item. More specifically, the original item difficulty parameter in the Rasch model is decomposed into the difficulty of the item and the effect of presenting the item in different positions. In case of a linear item position effect, the model reads as:

$$logit\left[Y_{pik}=1\right]=\theta_p-\beta_i+\left(\gamma+\delta_p\right)\left(k_{pi}-1\right), \tag{1}$$

where Y_{pik} is the response of person p to item i when presented in position k, and θ_p is the ability of person p measured given β_i, the difficulty of item i when presented at the first position of the test. k_{pi} is the position of item i that is presented to person p. γ is the fixed linear effect representing the general effect of item positions across persons, and δ_p is the random effect with $\delta_p \sim \left(0,\sigma_\delta^2\right)$ capturing individual deviation from the average γ for person p, referred to as persistence. Hence, the sum $\left(\gamma+\delta_p\right)$ denotes the individual changes in test performance across item positions, in which a positive value indicates an increase in performance, while a negative value indicates a decrease.

The model in Eq. 1 can be regarded as a generalization of the Rasch model or a logistic multilevel model with item responses nested within students. As PISA has a hierarchical data structure of students nested within schools, Debeer et al. (2014) extended the model in Eq. 1 to include the school level, resulting in a three-level model with item responses nested within students and students nested within schools:

$$logit\left[Y_{spik}=1\right]=\left(\theta_s+\theta_{ps}\right)-\beta_i+\left(\gamma+\delta_s+\delta_{ps}\right)\left(k_{pi}-1\right), \tag{2}$$

in which θ_s and θ_{ps} are the between-school part and within-school part for ability, respectively. A similar decomposition holds for the persistence parameters δ_s and δ_{ps}. The advantage of the extended model lies in the possibility of investigating whether the substantial variance in persistence is located at the school level or the individual level. It has been shown that variations of persistence between schools were rather small, on average 10% across countries in the PISA 2009 reading assessment (Debeer et al. 2014). However, in order to keep the statistical analyses aligned with the stratified sampling procedure of PISA and to filter out as much as possible "external" context factors (e.g.,

Table 1 Numbers of countries and students participating in the three PISA assessments, of which the major domains and numbers of items were included in the current study

PISA	Countries	Students	Major domain	Items included
2006	57	397,920	Science	103
2009	65	479,616	Reading	131
2012	68	485,490	Mathematics	109

school discipline, testing climate) that may contribute to individual persistence, the present study adopts the model in Eq. 2, maintaining the school level in the analyses.

Finally, to address the second research objective of exploring possible predictors of individual persistence, we further extend the previous model to include student characteristics as explanatory person covariates. For example, with a person level covariate Z_p, the model is extended as:

$$logit\left[Y_{spik}=1\right]=\left(\theta_s+\theta_{ps}\right)-\beta_i+\left(\gamma+\delta_s+\delta_{ps}\right)\left(k_{pi}-1\right)+\gamma_z Z_p+\gamma_{z\times position}Z_p\left(k_{pi}-1\right), \quad (3)$$

where γ_z is the fixed main effect of predictor Z_p on overall performance, and $\gamma_{z\times position}$ is the interaction effect between item positions and the predictor. The interaction effect captures the effect of predictor Z on individual persistence. Accordingly, δ_s and δ_{ps} are the remaining between-school and within-school variations of persistence that are not explained by predictor Z_p, respectively.

Two assumptions of these models should be noted. First, a linear item position effect on performance is assumed. Nonlinear functions, such as quadratic and cubic functions are also possible and have been evaluated in other studies (Debeer and Janssen 2013; Meyers et al. 2009). However, within the context of PISA assessments, Debeer and Janssen (2013) showed that the model with a random linear position effect provided the best fit in comparison with other model specifications. Hartig and Buchholz (2012) also calculated the percentages of correct responses in each position in PISA 2006 and found that they almost perfectly negatively correlated with the varying positions with more than 87% of the variance being explained by a linear trend. Given that there are only four possible varying positions in PISA assessments (see "Method" section below), a linear position effect seems to be reasonable and appropriate here. The second assumption of the models is that the position effect is constant for all items, since the interaction between item content and item position is not modeled in the current study.

Method

Data

The analyses made use of the published data from the PISA 2006, 2009, and 2012 paper-based assessments. Table 1 gives the overview of the numbers of countries and students participating in each of the three assessments.

The PISA test items consisted of a mixture of multiple-choice questions and constructed-response questions. Items of the same test domain were compiled into item clusters, and each test booklet was composed of four such item clusters. A balanced incomplete block design was employed in PISA so that each item cluster appeared once

Table 2 Cluster rotation design of PISA 2012

Booklet	Cluster position			
	1	2	3	4
1/21	PM5	*PS3*	PM6A/PM6B	*PS2*
2/22	*PS3*	*PR3*	PM7A/PM7B	*PR2*
3/23	*PR3*	PM6A/PM6B	*PS1*	PM3
4/24	PM6A/PM6B	PM7A/PM7B	*PR1*	PM4
5/25	PM7A/PM7B	*PS1*	PM1	PM5
6/26	PM1	PM2	*PR2*	PM6A/PM6B
7/27	PM2	*PS2*	PM3	PM7A/PM7B
8	*PS2*	*PR2*	PM4	*PS1*
9	*PR2*	PM3	PM5	*PR1*
10	PM3	PM4	*PS3*	PM1
11	PM4	PM5	*PR3*	PM2
12	*PS1*	*PR1*	PM2	*PS3*
13	*PR1*	PM1	*PS2*	*PR3*

Note PISA 2012 consisted of seven mathematics clusters (PM1–7; major domain), three science clusters (PS1–3), and three reading clusters (PR1–3). The minor domains are represented in italics. Booklets 1–13 were standard booklets. Booklets 21–27 were easier ones in which two standard mathematics clusters PM6A and PM7A were replaced by two easier ones PM6B and PM7B, respectively, and the other item clusters remained the same as in the standard booklets. The easier booklets were offered as an option to lower performing countries

and only once in each of the four possible cluster positions within a test booklet. In addition to the standard 2-h booklets, a special 1-h (UH) booklet, containing about half as many items as the standard ones, was prepared for schools catering for students with special needs. As including those students would jeopardize the random allocation of the balanced test design booklets, data from the UH booklets were excluded in the analyses.

For each PISA cycle, one of the reading, mathematics, or science subjects is chosen as the major domain and the other two remaining areas are minor domains. The major domain is assessed with more items, and each booklet contains at least one item cluster of the major domain. However, item clusters of the two minor domains are not always presented simultaneously in one test booklet, so some students, besides clusters of the major domain, only responded to items in either of the two minor domains. In such a case, the position effects of the items in minor domains would be considered as between-subjects, and may be confounded by the differences across groups of students who cause persistence. Therefore, to ensure more stable and reliable parameter estimates, only items of the major domains were used in the analyses, as shown in Table 1. The design of the item cluster rotation is illustrated in Table 2 using the PISA 2012 assessment. PISA 2006 and 2009 had the similar design. More detailed information about the paper-based assessment and test design can be found in the PISA technical reports (OECD 2009, 2012, 2014b).

Explanatory variables

After the cognitive assessment, a contextual questionnaire was administered to collect information on student characteristics, family background, and their attitudes towards the subject being assessed, school, and learning experiences. We made use of gender, the

socio-economic status index, language spoken at home, and three motivational indices from the PISA contextual questionnaires.

Socio-economic status (SES)

The PISA index of economic, social, and cultural status (ESCS) was derived from several indices including family wealth (e.g., "In your home, do you have a room of your own?"), home possessions of cultural and educational resources (e.g., "How many books are there in your home?"), the highest occupational status of parents, and the highest educational level of parents. The ESCS scores were obtained as component scores for the first principal component, with zero being the score of an average OECD student and one being the standard deviation across equally weighted OECD countries.

Enjoyment and interests

The scale of general enjoyment of and interest in the subject being assessed can be considered as a measure of the domain-specific motivation that depicts a relatively stable personal trait of the student (Penk et al. 2014). The joy of science (JOYSCIE) in PISA 2006 and the joy of reading index (JOYREAD) in PISA 2009 were constructed through the scaling of 5 and 11 items measuring students' enjoyment of science and reading, respectively. Students' interest in mathematics (INTMAT) was derived from four items measuring students' mathematics interest in PISA 2012. Students were asked to indicate their levels of agreement with the given statements (e.g., "Reading is one of my favorite hobbies.") by choosing one of the four response categories: "strongly disagree", "disagree", "agree", and "strongly agree".

Effort thermometer

The effort thermometer can be regarded as a measure of situation-specific motivation that describes a state of students depending on a specific situation (Baumert and Demmrich 2001; Penk et al. 2014). It was used in PISA 2006 and 2012, and was based on three 10-point scales collecting the information on students' motivation when completing the PISA assessment. Students were asked to indicate their effort expenditure (a) in a situation of high personal importance, (b) during the current PISA assessment, and (c) if the PISA test marks were to be counted in their school marks. Only (b) the effort spent during the current PISA assessment (EFFORT-REAL) was used in the present analyses. Note that the effort thermometer was not available in PISA 2009.

Perseverance (in mathematics)

The perseverance index (PERSEV) was constructed through the scaling of five items measuring students' sustained effort in solving a problem (e.g., "I continue working on tasks until everything is perfect."). Students were asked to respond by choosing one of five response categories: "very much like me", "mostly like me", "somewhat like me", "not much like me", and "not at all like me". This newly constructed index was only available in PISA 2012.

In PISA, the categorical items from the context questionnaires were scaled using IRT modeling, and the logits for the latent dimensions were transformed to scales with an OECD average of 0 and a standard deviation of 1 (with equally weighted samples).

The exact scaling procedure of the above scales can be found in PISA technical reports (OECD 2009, 2012, 2014b). The covariates used in the current study were in general uncorrelated or weakly correlated, with the correlations ranging from .004 between SES and perseverance in mathematics in PISA 2012 to .344 between interest in mathematics and perseverance in PISA 2012 across countries.

Data analysis[1]

Data coding

PISA used both dichotomous and partial credit scoring. In order to fit the dichotomous IRT models in Eqs. 2 and 3, partial credit items were dichotomized by scoring the full credit as correct (1) and all partial credits as incorrect (0). As a result, 6, 7, and 11 partial credit items were recoded in PISA 2006 science, 2009 reading, and 2012 mathematics, respectively. Following the calibration procedures in PISA (OECD 2009, 2012, 2014b), not-reached items were dropped as not administered, and items with missing responses were considered as incorrect (0).[2]

Gender was dummy coded with boys as the reference category. Language spoken at home was dummy coded with students speaking a different language from the language of the test as mother tongue as the reference group. Each item cluster can appear in different positions varying from Position 1 to 4 (Table 2), and within each cluster the positions of the items were fixed. Thus, the position of items in a test booklet k_{pi} was represented by the position of the corresponding clusters.

Analysis scheme

Because of the large sample size, the analyses were conducted separately for each country for each of the three PISA assessments. All analyses were conducted using the *glmer* function in the R package *lme4* (Bates et al. 2015) with a faster estimation procedure using an argument "nAGQ = 0".[3] The standard errors of estimates were computed using the default estimation in the *glmer* function. Although stratification variables can substantially reduce the standard errors of the estimates, they were not included in the computation of the current analyses so as to align with the previous studies whose results the current study aimed to replicate.

Table 3 gives an overview of the analysis scheme. The model in Eq. 2, with responses nested within students and students nested within schools, was used to address Research Question (RQ) 1a of examining the general position effect and 1b of individual persistence in PISA assessments (Model 0). The correlations of the cluster position effects across domains were calculated to see whether the effects were consistent within

[1] Given that in PISA, item fit is checked and badly behaving items are removed from the analysis, and because PISA tests sufficiently fit the Rasch model (e.g., OECD 2014b), we assume that a more complex model should fit the data at least equally well. Therefore, the model and item fit are not discussed in the results.

[2] The impacts of different treatments of missing responses on position effects are discussed in "Limitations" section.

[3] The "nAGQ" argument controls the number of points per axis for evaluating adaptive Gauss-Hermite quadrature approximation to log-likelihood. When nAGQ = 1 (default) indicates Laplace approximation. When nAGQ = 0, the random effects are not integrated out. The random effects and the fixed-effect coefficients are optimized in the penalized iteratively reweighted least squares step and estimation thus becomes faster (Bates et al. 2015). Doing so will have an impact on the estimation, but because of the large sample sizes in each country of PISA, the differences between nAGQ = 0 and nAGQ > 0 will be rather small, if not negligible. In fact, preliminary analyses with both nAGQ = 0 and nAGQ = 1 for 10 countries showed that the results were quite similar for the two estimation procedures. And because it was considerably faster, the estimation procedure with nAGQ = 0 was therefore chosen.

Table 3 Overview of the analysis scheme

Model	2006 science	2009 reading	2012 mathematics
0	No covariates	No covariates	No covariates
1	Gender SES Language at home	Gender SES Language at home	Gender SES Language at home
2	Joy of science	Joy of reading	Mathematics interest
3	Joy of science Effort	–	Mathematics interest Effort Perseverance

countries. The intraclass correlation (ICC) for persistence ICC_δ, which gives the proportion of variance located at the school level, was computed for each country. The correlations between ability and persistence (RQ 1c) were examined at the school level $\rho_{\theta_s\delta_s}$ and the student level $\rho_{\theta_{ps}\delta_{ps}}$.

Additionally, using the national mean scores reported by PISA in the corresponding domains as the country level performance, the correlations were calculated between the national mean scores and the estimates of cluster position effect γ, variance in persistence $\left(\sigma^2_{\delta s} + \sigma^2_{\delta ps}\right)$, as well as the correlations between ability and persistence between schools $\rho_{\theta_s\delta_s}$ and within-schools $\rho_{\theta_{ps}\delta_{ps}}$, to examine the cluster position effects and persistence at the country level.

Next, student characteristics and motivational scales were added to the model in Eq. 3 as student level predictors. Model 1 addressed the RQ 2a, examining the effects of student background characteristics of gender, SES, and language spoken at home. RQ 2b was approached in two steps because some test-taking effort-related variables were not available in all three PISA domains. Model 2 focused on the effects of the general enjoyment of and interest in the subject being assessed as the domain-specific motivation, and Model 3 further included the effort thermometer as the situation-specific motivation in PISA 2006 science and 2012 mathematics, and perseverance in 2012 mathematics.

Model parameters

Given the extensive results from different sets of analyses, we focused on the general item cluster position effect γ, persistence within-schools $\sigma^2_{\delta ps}$, and the fixed effects of student covariates on performance γ_z. Moreover, the most interesting parameters for the present study were the interactions between the explaining variables and the item cluster position, $\gamma_{z\times position}$, as they represent the effects of individual person covariates on persistence.

Since the analyses were conducted for each country separately, the obtained estimates were not necessarily on the same scale. To enable comparisons across countries, all the estimates were standardized using the standard deviation of the ability level within each country $\sqrt{\sigma^2_{\theta s} + \sigma^2_{\theta ps}}$ (cf., Debeer et al. 2014). For example, the position effect γ was standardized into $\gamma^* = \gamma / \sqrt{\sigma^2_{\theta s} + \sigma^2_{\theta ps}}$, presenting the average position effect on performance expressed in the standard deviation of ability within a country. The standard deviation of persistence between-school and within-school were standardized into $\sigma^*_{\delta s} = \sigma_{\delta s} / \sqrt{\sigma^2_{\theta s} + \sigma^2_{\theta ps}}$ and $\sigma^*_{\delta ps} = \sigma_{\delta ps} / \sqrt{\sigma^2_{\theta s} + \sigma^2_{\theta ps}}$, respectively. This resulted in a

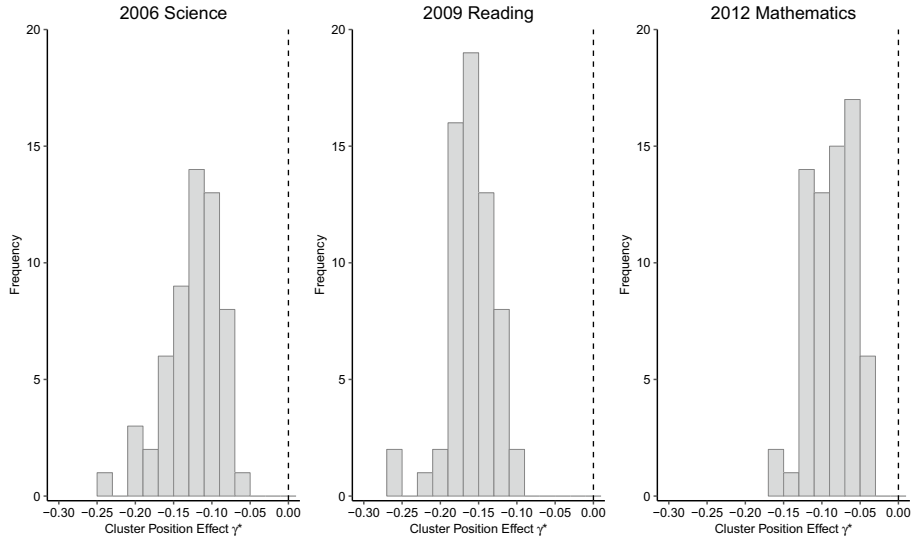

Fig. 1 The histograms of cluster position effects γ^* of all countries in the three PISA domains. The dotted lines indicate the value of zero. All effects are significantly different from zero at an alpha level of .05

standardized total variance in persistence $\left(\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}\right)$. The same standardization was also applied to the estimates of person covariates. In this way, the coefficients were transformed onto a common scale, representing the effect of one cluster position on performance relative to the standard deviation of the ability level within each country.[4]

Results

General effects of item cluster positions

The results of Model 0 showed that a significant negative effect of item cluster positions on overall performance was found in all countries in all three PISA domains. It means that students' probability of giving a correct response to an item decreased when the item was administered toward the end of the test. Figure 1 presents the distributions of the general cluster position effects relative to the standard deviation of the ability level within each country, γ^*, of all countries in the three domains. It can be seen that the strength of the cluster position effects varied considerably across countries. For instance, in PISA 2012 mathematics the effects ranged from $-.162$ in Mexico to $-.039$ in Vietnam. Note that the effect γ^* indicates the effect of one cluster position, and it would be three times larger if an item is presented at the end of the booklet (i.e., three cluster positions further). In terms of the probability of correct responses, this could result in a decrease of .040 in Mexico and .001 in Vietnam for a student with average ability ($\theta = 0$) when an item of average difficulty ($\beta_i = 0$) was administered one cluster further. When the item was placed at the end of the test, the decrease in the success probability would be rather substantial, varying from .119 in Mexico to .029 in Vietnam.

[4] The original parameters are available on request.

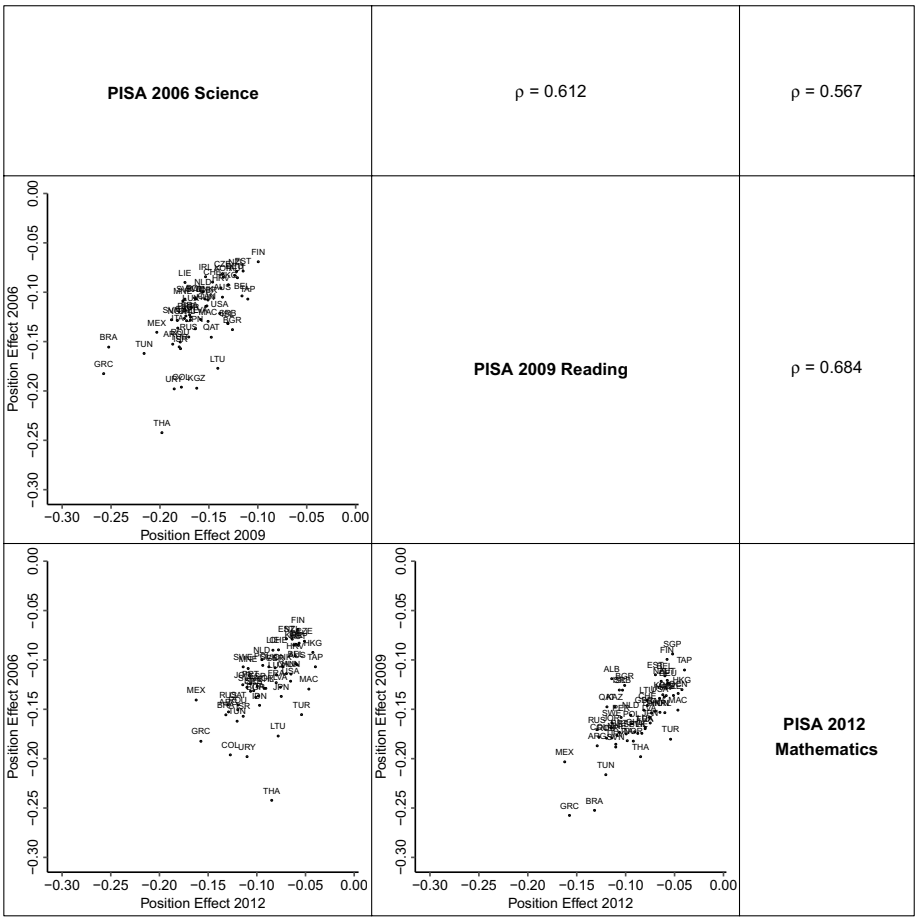

Fig. 2 The scatter plots and the correlations of cluster position effects of all countries across the three PISA domains

Figure 1 also shows that the cluster position effects were on average smaller in mathematics than those in science and reading. Figure 2 presents the scatter plots and the correlations of the cluster position effects γ^* of all countries across the three domains. It can be seen that the effects were considerably stable across PISA domains at the country level, suggesting that countries with a strong position effect in, for example, reading also tended to have strong effects in science and mathematics.

Persistence

Although numerically small, significant variances of persistence were found in all countries in the three PISA assessments, both between schools $\sigma^2_{\delta s}$ and within schools $\sigma^2_{\delta ps}$. Table 4 gives the mean ICC for persistence across countries in each domain. It shows that only a small proportion of variance of persistence can be explained by between-school differences, with about 6.3%–8.5% on average across countries. The substantial part of variability in persistence was attributed to individual differences across students.

Table 4 also gives the average correlations between persistence and proficiency at the school and the student levels across countries. Between schools, positive correlations between persistence and ability were found, meaning that schools with higher abilities

Table 4 Mean ICC for persistence, mean correlations of ability and persistence between schools and within schools, and standard deviations across countries in PISA assessments

PISA cycle and domain	$\overline{ICC}_{\delta}$ (*SD*)	$\overline{\rho}_{\theta_s\delta_s}$ (*SD*)	$\overline{\rho}_{\theta_{ps}\delta_{ps}}$ (*SD*)
2006 science	.079 (.062)	.428 (.620)	−.203 (.204)
2009 reading	.085 (.044)	.427 (.345)	−.132 (.125)
2012 mathematics	.063 (.052)	.225 (.507)	−.189 (.144)

Note ICC_{δ}, the intraclass correlation for persistence; $\rho_{\theta_s\delta s}$, correlation between persistence and ability at the school level; $\rho_{\theta_{ps}\delta ps}$, correlation between persistence and ability at the student level

Table 5 Correlations of PISA national mean scores with cluster position effects, total variance in persistence, and between-school and within-school correlations between ability and persistence in the three PISA assessments

Estimated parameter	PISA national mean score		
	2006 science	2009 reading	2012 mathematics
Cluster position effect γ^*	.695	.340	.680
Total variance in persistence $\left(\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}\right)$	−.303	−.624	−.507
Between-school correlation $\rho_{\theta_s\delta_s}$	.593	.580	.704
Within-school correlation $\rho_{\theta_{ps}\delta_{ps}}$	.446	.621	.662

Note All correlations are significant at an alpha level of .05

tended to have a higher level of persistence. Moreover, the correlations varied largely across countries with large standard deviations. It needs to be kept in mind that the variance in persistence between schools was very small. Within schools, on the other hand, there were negative correlations between persistence and ability, although these correlations were rather weak. This suggests that students with a higher proficiency level tended to have a slightly lower level of persistence, i.e., having a larger decrease in their performance. It could be possible that the motivation of higher ability students was high in the beginning and dropped as the test went along, whereas lower ability students had a rather low level of motivation to begin with.

Table 5 presents the correlations of the national performance level in PISA with the cluster position effect γ^*, the total variance in persistence $\left(\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}\right)$, and the correlations between ability and persistence at the school level and at the student level for the three PISA assessments. The cluster position effects were positively correlated with the PISA national scores,[5] ranging from a medium correlation of .340 in 2009 reading to stronger ones in 2006 science and 2012 mathematics (.695 and .680, respectively). These correlations suggest that the negative cluster position effects were more pronounced in lower performing countries. The negative correlations between the national mean scores and the total variance in persistence indicate that there was also more variability in

[5] The correlation of the PISA country scores with persistence may be confounded, since the PISA scores may be affected by persistence. However, it is assumed that the effects of persistence on the overall performance level to be very small relative to the between-country variation in overall performance. In theory, the correlation between country ability level and persistence could be estimated in a four-level model (responses in students in schools in countries), but this is not feasible in practice due to the large size of the data set (see "Limitations" section).

persistence in lower performing countries than in higher performing countries. Furthermore, the within-school correlations between persistence and ability were higher (closer to zero) in higher performing countries, meaning that persistence was more likely to be uncorrelated with ability in those countries.

Effects of person covariates

Main effects

Figure 3 presents an overview of the main effects of student predictors on overall performance. These effects essentially replicated the findings reported in PISA result reports (e.g., OECD 2007, 2010, 2014a). Girls were associated with a higher level of performance in reading, but a lower proficiency level in science and mathematics. A higher SES, speaking the same language of the test at home, having interest in the subject, a higher level of test-taking effort, and a higher level of perseverance (in mathematics) had positive effects on the test performance. It is interesting to note that the sizes of the effect of speaking the test language at home were generally several times larger than those of the other predictors, even after controlling for the SES of students. A possible explanation can be that all the assessments, though different in domain contents, require reading and comprehension of questions. Consequently, a good mastery of the language of the test can have a bigger effect on the achievement on the assessments.

Interaction effects

Figure 4 gives an overview of the distributions of the effects of student covariates on persistence of all countries in the three PISA domains. Overall, the effect sizes were relatively small (below .1) and no clear patterns can be discerned, except for gender. Girls tended to have a higher level of persistence, meaning that the decrease in performance related to the item position effect was smaller for girls than for boys. SES and speaking the test language at home did not show significant effects on persistence in the majority of the countries. Enjoyment of doing the subject as domain-specific motivation showed varying effects on persistence across subject domains. Only in PISA 2009, enjoyment of reading had a positive effect in most countries, in which students who enjoyed reading tended to have higher persistence than those who did not. This positive relationship was, however, not observed in 2006 science and 2012 mathematics in most of the countries. Furthermore, students' self-reported test-taking effort showed a positive effect on persistence in 2006 science and 2012 mathematics in some but not all participating countries, suggesting that a higher level of test-taking motivation could, to some extent, help to maintain the test performance levels. Students' self-reported perseverance did not show a significant effect on persistence in PISA 2012 mathematics in most countries, after controlling for the test-taking effort.

Discussion

Item position effects have been repeatedly shown in research with large-scale achievement assessments. Within this paper, a multilevel IRT model and its extensions were implemented to provide a systematic examination of item position effects in the three major domains of PISA assessments. The results replicated and extended the findings reported in the 2006 science (Hartig and Buchholz 2012) and 2009 reading assessments

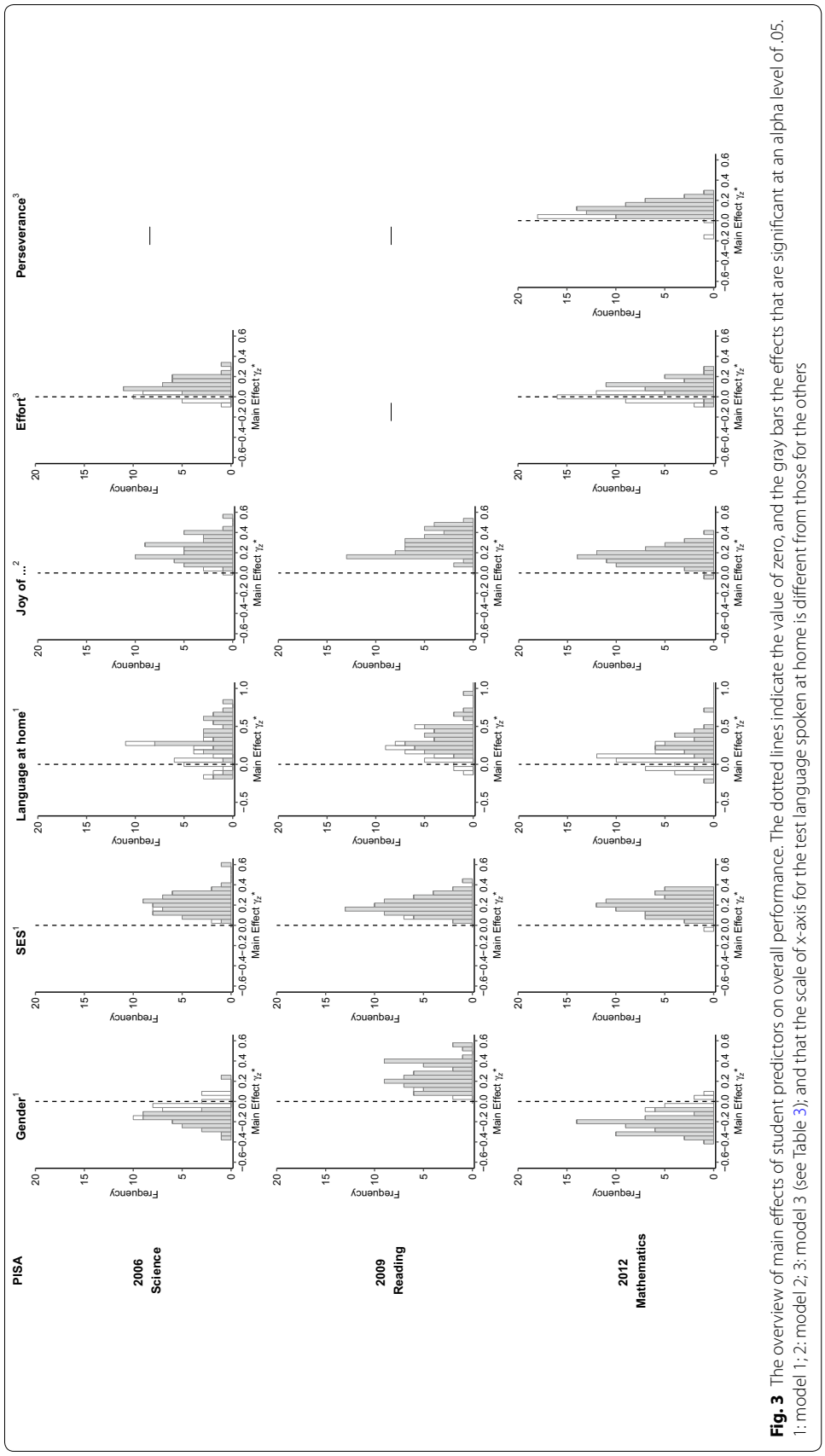

Fig. 3 The overview of main effects of student predictors on overall performance. The dotted lines indicate the value of zero, and the gray bars the effects that are significant at an alpha level of .05. 1: model 1; 2: model 2; 3: model 3 (see Table 3); and that the scale of x-axis for the test language spoken at home is different from those for the others

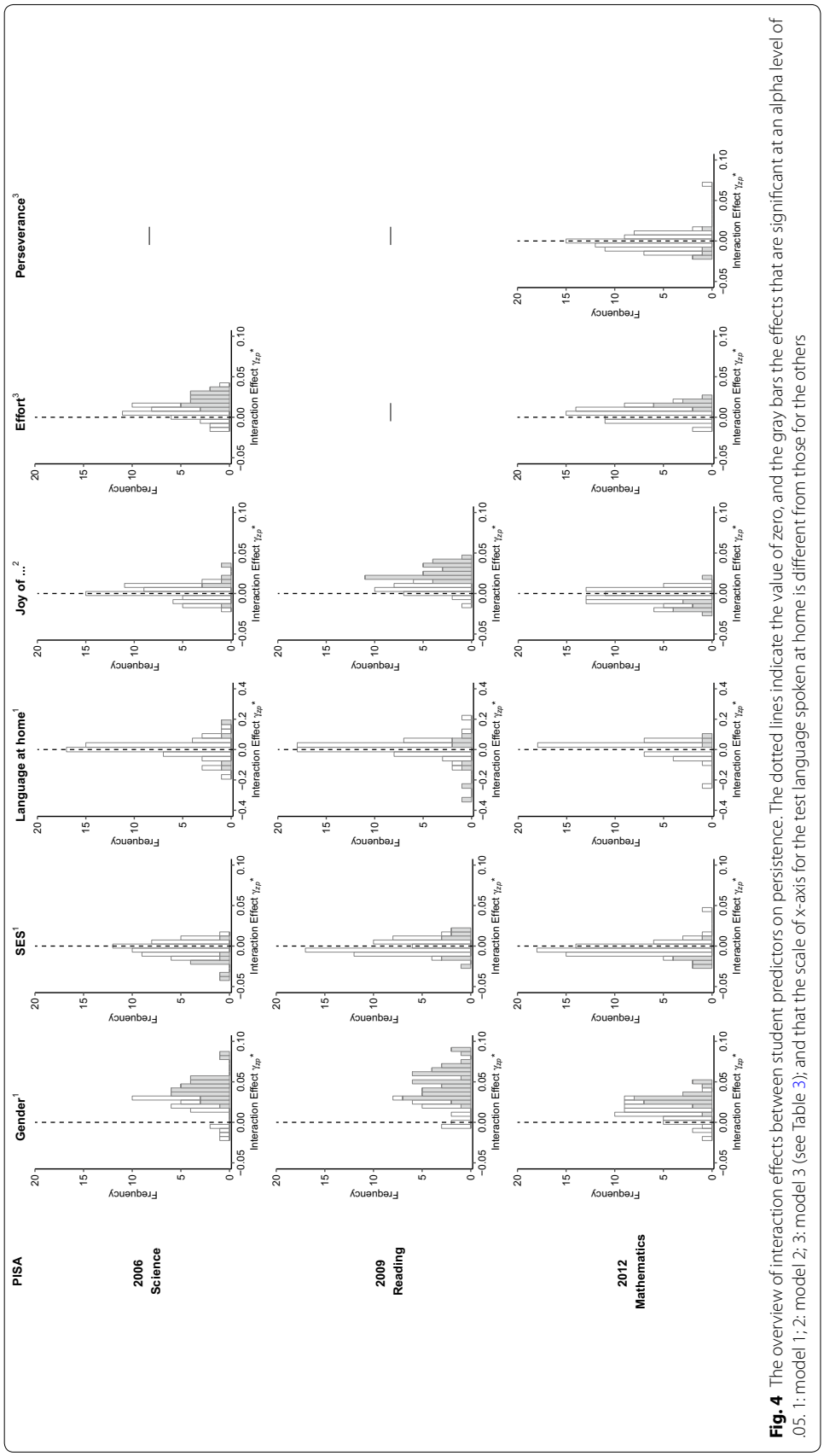

Fig. 4 The overview of interaction effects between student predictors on persistence. The dotted lines indicate the value of zero, and the gray bars the effects that are significant at an alpha level of .05. 1: model 1; 2: model 2; 3: model 3 (see Table 3); and that the scale of x-axis for the test language spoken at home is different from those for the others

(Debeer et al. 2014), and also showed a consistent pattern of a negative position effect and significant individual differences in the 2012 mathematics assessment. Furthermore, the analyses were extended to provide an exploratory investigation of possible predictors of individual differences in position effects, which may shed some light on why item position effects occur and on the nature of individual persistence.

The general negative effects of item cluster positions on performance found in our study confirm the findings in previous studies on item position effects (e.g., Debeer et al. 2014; Hartig and Buchholz 2012; Kingston and Dorans 1984; Meyers et al. 2009; Weirich et al. 2016). The effects were stronger in lower performing countries, despite the fact that some of them were administered easier booklets. On the other hand, it is also possible that the lower performance in those countries was due to their students being more affected by the cluster position effect. Hence, it is not unlikely that the PISA score differences between lower and higher performing countries could be reduced to some extent if the item cluster position effects had been taken into account. Moreover, the position effects were rather stable across the PISA domains at the country level, that is, lower performing countries with a strong position effect in reading also tended to demonstrate strong effects in science and mathematics.

The effects due to different item cluster positions as such may not invalidate comparisons between persons if every test taker would be affected by the effects in the very same way. However, in line with the findings of previous studies (Debeer et al. 2014; Hartig and Buchholz 2012; Weirich et al. 2016; Wise et al. 1989), we demonstrated that the strength of the cluster position effects varied considerably across individuals and that these individual differences were more pronounced in lower performing countries. In particular, in higher performing countries the correlation between performance level of students and persistence was close to zero, suggesting the two constructs were practically uncorrelated. However, the negative correlation between persistence and the performance level was much stronger in lower performing countries, suggesting that to some extent these countries' lower PISA results may be due to the fact that students with a higher performance level were not able to sustain their performance levels during the test session.

When exploring possible predictors of persistence, we only observed a consistent effect of gender with girls being able to keep up their performance better than boys in all three domains. Previous research showed that girls are often associated with a higher level of motivation and effort during low-stakes assessments (Butler and Adams 2007; Eklöf 2007; Kornhauser et al. 2014). As a negative item position effect can be caused by a decrease in motivation, our findings may therefore suggest that there are also possible gender differences in how students sustain their motivation during testing. A positive effect of domain-specific motivation (enjoyment of the subject) on persistence was, however, only found for reading. For science and mathematics, there may be some other situation-specific motivation-related factors that contribute substantially to individual persistence.

It needs to be noted that the context questionnaires were administered after the 2-h cognitive assessments in PISA. The validity and reliability of these self-reported scales may also be influenced by the decreasing motivation and effort of students at the end of the long testing sessions. Also, the motivational scales, i.e., enjoyment of the subject,

test-taking effort and perseverance, were self-reported measures. They may suffer from students' incapability of accurately reporting them and from response style bias (Finn 2015), which may to some extent contribute to the inconsistent patterns of the predictors on persistence in some countries in our study. Further examination of the possible response style bias in students' responses to the context questions across countries may be needed. In PISA 2015, a shift from the paper-based assessment to a computer-delivered assessment has been implemented across the vast majority of participating countries. This enables the measure of new and expanded aspects of the domain constructs (OECD 2017), and also allows to use response time as a measure of test-taking effort to investigate the effect of motivation on persistence for future studies (e.g., Setzer et al. 2013; Silm et al. 2013; Wise and Kong 2005).

To conclude, although PISA applies a balanced booklet test design, presenting item clusters in different positions within a booklet can still have an unintended impact on test performance, and this effect can vary considerably across individuals and across groups. However, before the question of "how to correct for item position effects" to be answered, the question of "whether to correct for item position effects" should be dealt with first. That is, whether item position effects would bias the ability estimation and pose a threat to the validity of assessments depends on whether the variation in test performance caused by item position effects is considered as construct-irrelevant variance. PISA defines the construct of interest to be students' "ability to use their knowledge and skills to meet real-life challenges" (e.g., OECD 2014b, p. 22). If being able to sustain the effort and keep solving problems for a duration of two hours is regarded as part of this competence construct, then effects related to item positions are construct-relevant. On the other hand, if the intended competence construct focuses on the ability of solving some distinct problems, the inter-individual differences in position effects as well as differences between groups may render comparisons between subgroups of test takers unfair and invalid as they are influenced by the individual and group differences in the construct-irrelevant effects (Messick 1995). In such a situation, it is important to model and control for the effects related to item positions using, for example, the presented IRT models.

Nevertheless, the present study explored item position effects from the perspective of individual differences, and showed that the negative position effect varied between genders and between higher- and lower-performing countries. These effects can also be evaluated from the perspective of item characteristics. It would be interesting for future studies to investigate the interactions of item format/content with item positions, looking for which types of items are more affected by position effects. In addition, the current study only focused on the PISA assessments. Potential future studies could try to replicate the findings using other large-scale assessment data, such as TIMSS and the Progress in International Reading Literacy Study (PIRLS).

Limitations

Several limitations of this study should be noted. First, because of the large sample size of the PISA data, the analyses were performed country by country. The estimates were expressed relative to individual differences in ability within a country to facilitate comparisons between countries. To enable the direct comparison between countries, the

models used in the study can be further extended to include the country level to analyze all the countries simultaneously. However, the huge total sample size would demand extensive computing power that far exceeds the capacity of most personal computers.

Second, it needs to be noted that the results in this study can be to some extent influenced by the treatment of missing responses. The current analyses followed the same procedure as PISA used for item calibration, in which the missing responses on omitted items were treated as incorrect. All consecutive missing responses (except the first one in the series) on items at the end of the test were treated as not administered (OECD 2009, 2012, 2014b). However, research (e.g., Debeer et al. 2017; Rose et al. 2010) has shown that missing responses due to omitted and/or not-reached items hardly occur randomly and should not be simply ignored or treated as incorrect ad hoc.

On the one hand, when a certain group of students did not reach the end of test due to lack of motivation, test-taking strategy, time constraint, and so on, treating not-reached items as not administered may affect the estimation of the item difficulty for items in the end positions. In order to check the impact of not-reached items on position effects, we selected three countries in PISA 2012 with high proportions of not-reached items (Colombia, Peru, and Uruguay), excluded all the students with not-reached responses, and re-analyzed the data with the reduced samples. The effects (see Table 6 in Appendix) were similar to the ones with the complete sample, which implies that the effects were not (largely) affected by the students who did not reach the end of the test. Moreover, if not-reached items would be treated as incorrect, position effects would probably increase due to test takers not reaching the end of the test.

On the other hand, missing responses due to omitted items are shown to be more associated with test takers' proficiency and item characteristics. For example, difficult items (Rose et al. 2010) and open-ended questions (Okumura 2014) were more likely to be skipped than easier items and multiple-choice questions, respectively. Hence, missing responses due to omitted items can be informative and should not be ignored. Yet, treating them as incorrect may also lead to biased person and item parameter estimates (e.g., Rose et al. 2017). As PISA did, missing responses due to omitted items were treated as incorrect in the current study. It is possible that the position effects found in our study were partly due to omitted items in later clusters, as students would skip items because of fatigue or decreased motivation, and this effect would be stronger for open-ended questions. But still, countries (e.g., Romania, Shanghai-China, and the Netherlands) with very low proportions of omitted responses in PISA 2012 also demonstrated a negative position effect. To see the effect of omitted items, we selected three countries (Albania, Argentina, and Montenegro) in PISA 2012 with high proportions of omitted items, treated the omitted responses as missing, and dropped them from the analyses. The negative position effects were still found, although less pronounced than the ones when the omitted responses were considered as incorrect (see Table 7 in Appendix). It suggests that the observed position effects can be indeed partially (but not completely) due to the omitted responses in item clusters in the later positions of the assessment. Although there have been several modeling approaches proposed, modeling missing responses is beyond the modeling framework of the current study. The effects of missing data on position effects can be the focus of further research.

Finally, PISA items were compiled into clusters and presented in one of the four cluster positions within a booklet. Ideally, the local dependence between items of the same theme (e.g., in reading items) within a cluster should be taken into account. However, the current study used the positions of the clusters as an approximation of the item positions in the analyses. It is the cluster position effects that were modeled, and hence the dependence between items within clusters should not affect the cluster position effects.

Authors' contributions
The study was conceptualized by all authors. DD, JB and JH, and QW and RJ conducted the data analyses of PISA 2006 science, 2009 reading, and 2012 mathematics assessments, respectively. QW prepared the draft of the manuscript, and all authors contributed to the revisions. All authors read and approved the final manuscript.

Author details
[1] KU Leuven, Leuven, Belgium. [2] University of Zurich, Zurich, Switzerland. [3] German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany.

Acknowledgements
The authors would like to thank the editor and two anonymous reviewers for their comments on the previous version of the manuscript.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
All data and materials are available on the PISA database on the website (http://www.oecd.org/pisa/data/).

Funding
Not applicable.

Appendix

Comparisons of the effects of different treatments of missing responses

See Tables 6, 7.

Table 6 The average number (n) of not-reached items and comparison of position effects in PISA 2012 mathematics between the complete sample and the reduced sample with students with not-reached item responses excluded in three selected countries

Country	*n*	Complete sample		Reduced sample	
		Cluster	*SE*	Cluster	*SE*
Colombia	5.04	−.138	.007	−.107	.008
Peru	5.80	−.122	.008	−.081	.010
Uruguay	4.59	−.136	.009	−.117	.010

Table 7 The Average number (n) of omitted items and comparison of position effects in PISA 2012 mathematics between the different treatments of nonresponses due to omission in three selected countries

Country	*n*	Omitted as incorrect		Omitted dropped	
		Cluster	*SE*	Cluster	*SE*
Albania	13.23	−.127	.009	−.065	.009
Argentina	9.99	−.142	.008	−.087	.008
Montenegro	11.37	−.121	.008	−.071	.009

Publisher's Note

Received: 31 October 2018 Accepted: 12 March 2019
Published online: 15 March 2019

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bates, D. M., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16,* 441–462. https://doi.org/10.1007/BF03173192.

Berezner, A., & Adams, R. J. (2017). Why large-scale assessments use scaling and item response theory. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 323–356). Chichester, United Kingdom: Wiley. https://doi.org/10.1002/9781118762462.ch13.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8,* 279–304.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-3990-9.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39,* 502–523. https://doi.org/10.3102/1076998614558485.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50,* 164–185. https://doi.org/10.1111/jedm.12009.

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement, 54,* 333–363. https://doi.org/10.1111/jedm.12147.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7,* 311–326. https://doi.org/10.1080/15305050701438074.

Finn, B. (2015). *Measuring motivation in low-stakes assessments*. (ETS Research Report Vol. 15–19). Retrieved from http://doi.wiley.com/10.1002/ets2.12067.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54,* 418–431.

Hohensinn, C., Kubunger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly, 50,* 391–402.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147–154. https://doi.org/10.1177/014662168400800202.

Kornhauser, Z., Minahan, J., Siedlecki, K., & Steedle, J. (2014). *A strategy for increasing Student motivation on low-stakes assessments*. Retrieved from Council for Aid to Education website: http://cae.org/images/uploads/pdf/A_Strategy_for_Increasing_Student_Motivation.pdf.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387–413. https://doi.org/10.3102/00346543055003387.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22,* 38–60. https://doi.org/10.1080/08957340802558342.

Mullis, I. V. S., Martin, M. O., & Diaconu, D. (2004). Item analysis and review. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report* (pp. 225–251). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

OECD. (2009). *PISA 2006 technical report*. Paris, France: OECD. https://doi.org/10.1787/9789264048096-en.

OECD. (2010). *PISA 2009 results: What students know and can do*. Paris, France: OECD. https://doi.org/10.1787/9789264091450-en.

OECD. (2012). *PISA 2009 technical report*. Paris, France: OECD. https://doi.org/10.1787/9789264167872-en.

OECD. (2014a). *PISA 2012 results: What students know and can do (Volume I) (Rev. ed)*. Paris, France: OECD. https://doi.org/10.1787/9789264208780-en.

OECD. (2014b). *PISA 2012 technical report*. Paris, France: OECD.

OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD.

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educational and Psychological Measurement, 74,* 611–626. https://doi.org/10.1177/0013164413516976.

OECD. (2007). *PISA 2006 science competencies for tomorrow's world: Volume 1: Analysis*. Paris, France: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/9789264040014-en.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education, 2*(5), 1–17. https://doi.org/10.1186/s40536-014-0005-4.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82,* 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Report No. RR-10-11). Princeton, NJ: Educational Testing Service.

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science, 51*, 47–64.

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50,* 1249–1254. https://doi.org/10.1016/j.paid.2011.02.019.

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26,* 34–49. https://doi.org/10.1080/08957347.2013.739453.

Silm, G., Must, O., & Täht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *Trames, 17,* 433–448. https://doi.org/10.3176/tr.2013.4.08.

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement, 24,* 151–162. https://doi.org/10.1177/01466210022031589.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2016). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*, 115–129. https://doi.org/10.1177/0146621616676791.

Wise, L. L., Chia, W. J., & Park, R. (1989). *Item position effects for test of word knowledge and arithmetic reasoning.* In Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18,* 163–183. https://doi.org/10.1207/s15324818ame1802_2.